การทดสอบข้อตกลงเบื้องต้นบางประการของสถิติ 12

ในการใช้สถิตินั้น ถ้าหากจะได้ผลการวิเคราะห์สามารถเชื่อถือได้สูง ค่าที่ได้ถูกต้องตรงตามความเป็นจริงแล้วล่ะก็ จำเป็นอย่างยิ่งในการเลือกใช้สถิติแต่ละตัว จะต้องคำนึงถึงข้อตกลงเบื้องต้นบางประการของสถิติเสียก่อน ซึ่งสถิติแต่ละตัวก็มีข้อตกลงเบื้องต้นแตกต่างกันไป ข้อตกลงเบื้องต้นบางประการของสถิติผู้วิจัยสามารถจัดกระทำได้โดยตรง เช่น สถิติพาราเมตริกซ์ทุกตัวจะมีข้อตกลงเบื้องต้นว่าการวัดจะต้องอยู่ในมาตราการวัดตั้งแต่ระดับอันตรภาคชั้น (Interval Scale) ขึ้นไป เป็นต้น แต่ข้อตกลงเบื้องต้นบางประการของสถิติผู้วิจัยไม่สามารถจะจัดกระทำได้ ต้องอาศัยการทดสอบทางสถิติเข้าช่วย เพื่อพิจารณาดูว่าข้อมูลที่ได้เป็นไปตามข้อตกลงเบื้องต้นหรือไม่

ข้อตกลงเบื้องต้นบางประการที่ต้องใช้การทดสอบทางสถิตเข้าช่วยมีดังนี้

1. คะแนนจะต้องมาจากประชากรที่มีการแจกแจงเป็นโค้งปกติ (Normal Distribution)

ข้อตกลงข้อนี้เป็นข้อตกลงเบื้องต้นในสถิติพาราเมตริกซ์ ลักษณะการแจกแจงของประชากรจะต้องเป็นโค้งปกติ เราสามารถทดสอบข้อมูลที่เก็บรวบรวมมาว่ามาจากประชากรที่มีการแจกแจงเป็นโค้งปกติหรือไม่โดยใช้สถิติไคสแควร์กลุ่มตัวอย่างเดียว (Chi-Square Goodness of Fit) หรือ The kolmogorow-Smirnov Test

สามารถตั้งสมมติฐานได้ว่า

H_0 : ข้อมูลมีการแจกแจงเป็นโค้งปกติ

H_1 : ข้อมูลมีการแจกแจงไม่เป็นโค้งปกติ

ดูวิธีการคำนวณในบทที่ 10 การทดสอบสถิติไร้พารามิเตอร์ในหัวข้อสถิติไคสแควร์กลุ่มตัวอย่างเดียว และ The Kolmogorow-Smirnov Test

2. ความเป็นเอกพันธ์ของความแปรปรวน (Homogeneity of Variance)

ข้อตกลงข้อนี้เป็นข้อตกลงเบื้องต้นในสถิติพาราเมตริกซ์ ประชากรทุกกลุ่มที่ศึกษาจะต้องมีการความแปรปรวนเท่ากัน

การทดสอบนั้น มีสถิติที่ใช้อยู่หลายสูตรด้วยกัน ขึ้นอยู่กับจำนวนของกลุ่ม

2.1 กรณีกลุ่มตัวอย่างสองกลุ่ม

ใช้สูตร $F = \frac{S_1^2}{S_2^2}$

$df_1 = n_1 - 1$ และ $df_2 = n_2 - 1$

ตั้งสมมติฐานได้ว่า

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

ในการใช้โปรแกรม SPSS for Windows ให้ใช้การทดสอบ t-test Independent ทดสอบความแปรปรวนของสองกลุ่มตัวอย่าง ดูในบทที่ 4

2.2 กรณีหลายกลุ่มตัวอย่าง

การทดสอบความแปรปรวนหลายกลุ่มตัวอย่าง มีวิธีการทดสอบคือ Bartlett Box F, Cochran' C และ Hartley's F max

ตั้งสมมติฐานได้ว่า

$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$

H_1 : มีความแปรปรวนอย่างน้อย 1 คู่ที่ไม่เท่ากัน

ในการใช้โปรแกรม SPSS for Windows ให้ใช้การทดสอบความแปรปรวนแบบทิศทางเดียว (One-way ANOVA) ดูในบทที่ 5

3. ความสัมพันธ์เชิงเส้นตรง (Linearity)

เป็นข้อตกลงในสถิติที่เกี่ยวกับความสัมพันธ์ระหว่างตัวแปร เช่น การวิเคราะห์การถดถอยและการวิเคราะห์สหสัมพันธ์ ว่าตัวแปรอิสระและตัวแปรตามที่จะวิเคราะห์นั้น จะต้องมีความสัมพันธ์กันเชิงเส้นตรง (Linearity)

ตั้งสมมติฐานได้ว่า

H_0 : ตัวแปรทั้งสองมีความสัมพันธ์กันเชิงเส้นโค้ง

H_1 : ตัวแปรทั้งสองมีความสัมพันธ์กันเชิงเส้นตรง

ตัวอย่าง 12.1 ตัวแปร X และ Y มีข้อมูลดังนี้

X	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
Y	2	3	4	3	4	5	3	4	5	4	5	6	4	5	6

ใช้คำสั่ง “Analyze” เมนูรอง “Compare Means…” และเมนูย่อย “Means…” คลิกเลือกตัวแปร X คือตัวแปรอิสระ ใส่ในช่อง “Independent List :” และตัวแปร Y คือตัวแปรตาม ใส่ในช่อง “Dependent List :” คลิกปุ่ม “Options…” คลิกเลือกที่ “Test for linearity”

ผลการวิเคราะห์มีดังนี้

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
x * y	Between Groups	(Combined)	12.500	4	3.125	1.786	.208
		Linearity	12.228	1	12.228	6.988	.025
		Deviation from Linearity	.272	3	.091	.052	.984
	Within Groups		17.500	10	1.750		
	Total		30.000	14			

Measures of Association

	R	R Squared	Eta	Eta Squared
x * y	.638	.408	.645	.417

ในการทดสอบ Linearity มีค่า F-test 6.988 มีนัยสำคัญทางสถิติ แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันเชิงเส้นตรง และ Deviation from Linearity มีค่า F-test .052 ไม่มีนัยสำคัญทางสถิติ นั่นคือ ตัวแปรทั้งสองไม่มีความสัมพันธ์กันเชิงเส้นโค้ง

สำหรับค่า Eta นั่นเป็นค่าสหสัมพันธ์กรณีความสัมพันธ์ระหว่างตัวแปรทั้งสองเป็นเส้นโค้ง และค่า Eta Square แปลความหมายเช่นเดียวกับค่า R Square หรือก็คือสัมประสิทธิ์การอธิบาย เป็นการบ่งบอกถึงความสามารถของตัวแปร X สามารถอธิบาย Y ได้เท่าไหร่

4. ตัวแปรอิสระแต่ละตัวต้องไม่มีความสัมพันธ์กัน

เป็นข้อตกลงเบื้องต้นในการวิเคราะห์การถดถอยพหุคูณนั้น เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตามตัวหนึ่งกับตัวแปรอิสระหลาย ๆ ตัว ซึ่งการวิเคราะห์นี้มีข้อตกลงข้อหนึ่งว่าตัวแปรอิสระเหล่านี้จะต้องมีไม่มีความสัมพันธ์กัน หรือหากสัมพันธ์กันก็จะต้องมีความสัมพันธ์กันไม่สูงมากนัก แต่ในทางปฏิบัติบางครั้งจะพบว่าตัวแปรอิสระมีความสัมพันธ์กันสูง ในกรณีที่ตัวแปรอิสระเพียง 2 ตัวมีความสัมพันธ์กันสูงจะเรียกว่า ***Collinearity*** และในกรณีที่ตัวแปรอิสระมากกว่า 2 ตัว มีความสัมพันธ์กันสูงจะเรียกว่า ***Multicollinearity***

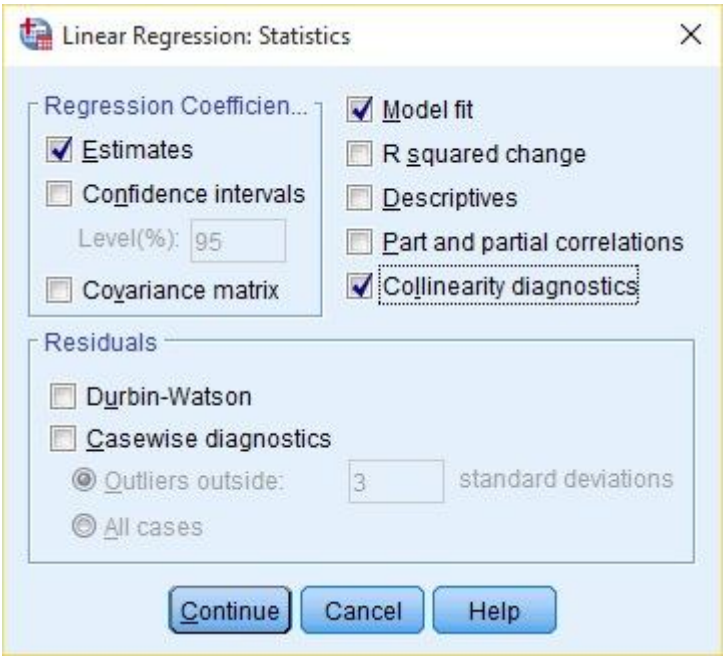
ตัวอย่าง 12.2 ในการวิเคราะห์การถดถอยของตัวแปรเกณฑ์คือ อัตราดอกเบี้ยที่ได้รับต่อปี (DIVI) บนตัวแปรอิสระ 2 ตัวคือ อัตราของรายได้ต่อปี (EARN) และจำนวนปี (TREND) ข้อมูลแสดงดังนี้

DIVI	2.80	3.16	3.40	3.70	4.10	4.50	5.00	5.00
EARN	4.13	4.63	5.17	5.80	6.21	6.83	7.35	7.91
TREND	1	2	3	4	5	6	7	8

สหสัมพันธ์ของตัวแปรทั้ง 3 คือ

Correlations:	DIVI (Y)	EARN (X_1)	TREND (X_2)
DIVI (Y)	1.0000	.9911**	.9923**
EARN (X_1)	.9911**	1.0000	.9996**
TREND (X_2)	.9923**	.9996**	1.0000

ในหน้าต่างของการวิเคราะห์การถดถอยจะมีปุ่มที่ชื่อว่า “Statistics…” คลิกที่ปุ่มนี้จะปรากฏหน้าต่าง “Linear Regression Statistics” จะมีเมนูที่ชื่อว่า “Collinearity diagnostics” ให้คลิกเลือกที่เมนูนี้ โปรแกรมจะประมวลผลแสดงค่าสถิติต่าง ๆ ที่เกี่ยวข้อง



ภาพประกอบ 12.1

มีดัชนีที่แสดงถึงปัญหา multicollinearity หลายตัวด้วยกัน ดังนี้

1. องค์ประกอบการขยายความแปรปรวน (Variance inflation factor : VIF)

VIF เป็นความสัมพันธ์ของตัวแปร X ตัวหนึ่งโดยการถดถอยบนตัวแปร X อื่น ๆ มีสูตรคำนวณคือ

$$VIF\ (X_i) \quad = \quad \frac{1}{1-R_i^2}$$

เมื่อ R_i^2 คือสัมประสิทธิ์ของการตัดสินใจ โดยการถดถอย X_i บนตัวตัวแปรอิสระอื่น ๆ ที่เหลือ

ถ้าตัวแปรอิสระทั้งหมดไม่สัมพันธ์กันแล้ว ค่า VIF จะมีค่าเป็น 1 ซึ่งค่า VIF โดยปกติจะมีพิสัยตั้งแต่ 1 ถึงอนันต์

เกณฑ์ในการพิจารณา VIF นั้น ขึ้นอยู่กับดุลยพินิจของผู้วิจัยอีกเช่นกัน แต่มีตำราบางเล่มเสนอแนะว่า ตัวแปรอิสระทั้งสองตัวจะเกิดปัญหา multicollinearity ก็ต่อเมื่อ ค่า VIF มีค่าตั้งแต่ 10 ขึ้นไป

2. Tolerence

ค่า tolerance สามารถคำนวณได้ด้วยสูตร

$$Tolerance \quad = \quad 1 - R^2 \quad = \quad \frac{1}{VIF}$$

ค่า tolerance มีค่าตั้งแต่ 0 ถึง 1 ถ้าหากค่าเข้าใกล้ 1 แสดงว่าตัวแปรเป็นอิสระจากกัน แต่ถ้าค่าเข้าใกล้ 0 แสดงว่าเกิดปัญหา multicollinearity

Variable	SE Beta	Correl	Part Cor	Partial	Tolerance	VIF
TREND	1.860104	.992333	.055953	.420980	8.4017E-04	1190.236
EARN	1.860104	.991128	-.027201	-.220092	8.4017E-04	1190.236

จะเห็นว่าค่า Tolerence มีค่า 0.0008401 มีค่าเข้าใกล้ 0 มาก ส่วนค่า VIF มีค่า 1190.236 จะเห็นว่าค่าทั้งสองเป็นไปตามเกณฑ์ของการเกิด multicollinearity

3. Condition Index และ Variance-Decomposition Proportions

ดัชนี 2 ตัวที่เราจะใช้ในการพิจารณาการเกิด multicollinearity ได้ก็คือ condition number (CN) และ condition index (CI) มีสูตรดังนี้

$$CN \quad = \quad \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

$$CI_i \quad = \quad \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

เมื่อ λ_{max} = ค่าไอเกนที่มากที่สุด (Largest eigenvalue) ; λ_{min} = ค่าไอเกนที่น้อยที่สุด (smalles eigenvalue) และ λ_{s} = ค่าไอเกนตัวที่ i

นอกจากนี้ยังต้องพิจารณาจาก Variance-Decomposition Proportions ซึ่งจะประกอบไปด้วยส่วนของจุดตัดและตัวแปรอิสระแต่ละตัว

Variance proportions ก็คือสัดส่วนของความแปรปรวนของจุดตัด (a) และสัมประสิทธิ์การถดถอย (b) แต่ละตัวที่สัมพันธ์กับ Condition index แต่ละตัว ซึ่งในแต่ละสดมภ์จะมีผลรวมเป็น 1.00 การแปลความหมายจะต้องนำ 100 ไปคูณ ดังเช่นในตัวแปร EARN จะแปลผลได้ว่า 0% ของความแปรปรวนใน b_{EARN} สัมพันธ์กับ Condition index ตัวแรก และ 0% สัมพันธ์กับ Condition index ตัวที่สอง และ 99.99% สัมพันธ์กับ Condition index ตัวที่สาม ส่วนสัมประสิทธิ์ b_{TREND} ก็แปลความได้เช่นเดียวกัน

Number	Eigenval	Cond Index	Variance Proportions		
			Constant	EARN	TREND
1	2.89046	1.000	.00001	.00000	.00002
2	.10952	5.137	.00042	.00000	.00086
3	.00002	359.497	.99957	.99999	.99912

การพิจารณา multicollinearity ในกรณีนี้มีตัวแปร 2 ตัวที่สัมพันธ์กัน ให้พิจารณาที่ Condition index ที่มีค่าสูงสุดอยู่ในบรรทัดสุดท้าย จากนั้นพิจารณาที่ variance proportion จะเห็นว่ามีค่าสูงพอ ๆ กันในตัวแปรทั้ง 2 ตัวที่สัมพันธ์กัน